NROM FLASH MEMORY WITH A HIGH-PERMITTIVITY GATE DIELECTRIC

TECHNICAL FIELD OF THE INVENTION

[0001] The present invention relates generally to memory devices and in particular the present invention relates to nitride read only memory (NROM) flash memory device architecture.

BACKGROUND OF THE INVENTION

[0002] Memory devices are typically provided as internal, semiconductor, integrated circuits in computers or other electronic devices. There are many different types of memory including random-access memory (RAM), read only memory (ROM), dynamic random access memory (DRAM), synchronous dynamic random access memory (SDRAM), and flash memory. One type of flash memory is a nitride read only memory (NROM). NROM has some of the characteristics of flash memory but does not require the special fabrication processes of flash memory. NROM integrated circuits can be implemented using a standard CMOS process.

[0003] Flash memory devices have developed into a popular source of non-volatile memory for a wide range of electronic applications. Flash memory devices typically use a one-transistor memory cell that allows for high memory densities, high reliability, and low power consumption. Common uses for flash memory include personal computers, personal digital assistants (PDAs), digital cameras, and cellular telephones. Program code and system data such as a basic input/output system (BIOS) are typically stored in flash memory devices for use in personal computer systems.

[0004] The performance of flash memory transistors needs to increase as the performance of computer systems increases. To accomplish a performance increase, the transistors can be reduced in size. This has the effect of increased speed with decreased power requirements.

[0005] However, a problem with decreased flash memory size is that flash memory cell technologies have some scaling limitations. For example, stress induced leakage typically requires a tunnel oxide above 60 Å. This thickness results in a scaling limit on the gate length. Additionally, this gate oxide thickness limits the read current and may require large gate widths.

[0006] For the reasons stated above, and for other reasons stated below which will become apparent to those skilled in the art upon reading and understanding the present specification, there is a need in the art for a more scalable, higher performance flash memory transistor.

SUMMARY

[0007] The above-mentioned problems with flash memory scaling and performance and other problems are addressed by the present invention and will be understood by reading and studying the following specification.

[0008] The present invention encompasses an NROM flash memory transistor with a high permittivity gate dielectric. The transistor is comprised of a substrate with a plurality of source/drain regions. The source/drain regions have a different conductivity than the substrate into which they are doped.

[0009] A high-k gate dielectric is formed on the substrate substantially between the plurality of source/drain regions. The gate dielectric has a high dielectric constant that is greater than silicon dioxide. The gate dielectric can be an atomic layer deposited and/or evaporated nanolaminate gate dielectric. A control gate is formed on top of the oxide insulator.

[0010] Further embodiments of the invention include methods and apparatus of varying scope.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] Figure 1 shows a cross-sectional view of one embodiment of an NROM flash memory cell transistor of the present invention.

[0012] Figure 2 shows an energy-band diagram in accordance with the transistor of Figure 1.

[0013] Figure 3 shows an energy-band diagram in accordance with a write operation to the transistor structure of Figure 1.

[0014] Figure 4 shows an energy-band diagram in accordance with an erase operation from the transistor structure of Figure 1.

[0015] Figure 5 shows a plot of tunneling current dependence on barrier height for various electric fields in accordance with the transistor structure of Figure 1.

[0016] Figure 6 shows a block diagram of an electronic system of the present invention.

DETAILED DESCRIPTION

[0017] In the following detailed description of the invention, reference is made to the accompanying drawings that form a part hereof and in which is shown, by way of illustration, specific embodiments in which the invention may be practiced. In the drawings, like numerals describe substantially similar components throughout the several views. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention. Other embodiments may be utilized and structural, logical, and electrical changes may be made without departing from the scope of the present invention. The following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims and equivalents thereof.

[0018] Figure 1 illustrates a cross-sectional view of one embodiment of a nitride read only memory (NROM) flash memory cell transistor of the present invention. This NROM transistor uses the high-k dielectric layer of the present invention as a trapping layer. In order to improve the programming speed and/or lower the programming voltage of an NROM device, it is desirable to use a trapping material with a lower conduction band edge (i.e., a higher electron affinity) to achieve a larger offset as well as to provide for programming by direct tunneling at low voltages.

[0019] High permittivity dielectric materials such as HfO_2 and ZrO_2 have a lower conduction band edge than the prior art trapping material, silicon nitride. If HfO_2 were used as a trapping layer, the offset would be 1.7 eV that is much better than the 1.2 eV associated with a nitride trapping layer.

[0020] The transistor is comprised of two source/drain regions 101 and 102 doped into the substrate 103. In one embodiment, these are n+ regions and the substrate is p-type silicon. However, the present invention is not limited to any conductivity type.

[0021] A tunnel oxide layer 105 is formed on the substrate 103 between the source/drain regions 101 and 102. The high dielectric constant trapping layer 107 is formed on top of the tunnel oxide layer 105 and another oxide layer 109 is formed on top of the trapping layer 107. The oxide – high-k dielectric – oxide layers 105, 107, and 109 form a composite gate insulator 100 under the polysilicon control gate 111. In one embodiment, the nanolaminate gate insulator 100 can be fabricated by the above-described ALD, the evaporated technique, or a combination of the two.

[0022] The simplest nanolaminates with high-k dielectrics are oxide – high-k dielectric – oxide composites. Since silicon dioxide has a low electron affinity and high conduction band offset with respect to the conduction band of silicon, these nanolaminates have a high barrier, Φ, between the high-k dielectric and the oxide. If the trapping center energies, E_t, in the high-k dielectrics illustrated in Figure 2 are large then other high-k dielectrics with a smaller barrier can be used.

[0023] Examples of oxide – high-k dielectric – oxide composites include: oxide – ALD HfO_2 – oxide, oxide – evaporated HfO_2 – oxide, oxide – ALD ZrO_2 – oxide, oxide – evaporated ZrO_2 – oxide, oxide – ALD ZrSnTiO – oxide, oxide – ALD ZrON – oxide, oxide – ALD ZrAlO – oxide, oxide – ALD $ZrTiO_4$ – oxide, oxide – ALD Al_2O_3 – oxide, oxide – ALD La_2O_3 – oxide, oxide – $LaAlO_3$ – oxide, oxide – evaporated $LaAlO_3$ – oxide, oxide – ALD $HfAlO_3$ – oxide, oxide – ALD HfSiON – oxide, oxide – evaporated Y_2O_3 – oxide, oxide – evaporated Gd_2O – oxide, oxide – ALD Ta_2O_5 – oxide, oxide – ALD TiO_2 – oxide, oxide – evaporated TiO_2 – oxide, oxide – ALD PrO_3 – oxide, oxide – evaporated PrO_3 – oxide, oxide – evaporated $CrTiO_3$ – oxide, and oxide – evaporated YSiO – oxide.

[0024] Another class of nanolaminates avoids tunneling between the trapping centers in the nitride layer of a conventional NROM device and the control gate. High-k dielectrics, in one embodiment, can be used as the top layer in the gate insulator nanolaminate. Since they have a much higher dielectric constant than silicon oxide, these layers can be much thicker and still have the same capacitance. The thicker layers avoid tunneling to the control gate that is an exponential function of electric fields but have an equivalent oxide thickness that is much smaller than their physical thickness.

[0025] Examples of this second category of nanolaminates include: oxide – nitride – ALD Al_2O_3, oxide – nitride – ALD HfO_2, and oxide – nitride – ALD ZrO_2.

[0026] A third category of nanolaminates employs traps in the high-k dielectrics that have a larger energy depth with respect to the conduction band in the high-k trapping layer than those in the first category. As a result, large offsets are not required between the layers in the nanolaminates and a wide variety of different nanolaminates are possible using only high-k dielectrics in these nanolaminates. The energy depths of the traps can be adjusted by varying process conditions.

[0027] Examples of this third category of nanolaminates include: ALD HfO_2 – ALD Ta_2O_5 – ALD HfO_2, ALD La_2O_3 – ALD HfO_2 – ALD La_2O_3, ALD HfO_2 – ALD ZrO_2 – ALD HfO_2, ALD Lanthanide (Pr, Ne, Sm, Gd, and Dy) Oxide – ALD ZrO_2 – ALD Lanthanide (Pr, Ne, Sm, Gd, and Dy) Oxide, ALD Lanthanide Oxide – ALD HfO_2 – ALD Lanthanide Oxide, and ALD Lanthanide Oxide – evaporated HfO_2 – ALD Lanthanide Oxide.

[0028] In one embodiment, the high-k gate dielectric layer is fabricated using atomic layer deposition (ALD). As is well known in the art, ALD is based on the sequential deposition of individual monolayers or fractions of a monolayer in a well controlled manner. Gaseous precursors are introduced one at a time to the substrate surface and between the pulses the reactor is purged with an inert gas or evacuated.

[0029] In the first reaction step, the precursor is saturatively chemisorbed at the substrate surface and during subsequent purging the precursor is removed from the reactor. In the second step, another precursor is introduced on the substrate and the desired films growth reaction takes place. After that reaction, byproducts and the precursor excess are purged from the reactor. When the precursor chemistry is favorable, one ALD cycle can be performed in less than one second in a properly designed flow-type reactor.

[0030] ALD is well suited for deposition of high-k dielectrics such as AlO_x, $LaAlO_3$, $HfAlO_3$, Pr_2O_3, Lanthanide-doped TiO_x, HfSiON, Zr-Sn-Ti-O films using $TiCl_4$ or TiI_4, ZrON, HfO_2/Hf, $ZrAlXO_y$, $CrTiO_3$, and $ZrTiO_4$.

[0031] The most commonly used oxygen source materials for ALD are water, hydrogen peroxide, and ozone. Alcohols, oxygen and nitrous oxide have also been used. Of these, oxygen reacts very poorly at temperatures below 600 ° C but the other oxygen sources are highly reactive with most of the metal compounds listed above.

[0032] Source materials for the above-listed metals include: zirconium tetrachloride ($ZrCl_4$) for the Zr film, titanium tetraisopropoxide ($Ti(OCH(CH_3)_2)_4$) for the Ti film, trimethyl aluminum ($Al(CH_3)_3$) for the Al film, chromyl chromide (CrO_2Cl_2) for the Cr film, praseodymium chloride ($PrCl_3$) for the Pr film, and hafnium chloride ($HfCl_4$) for the Hf film. Alternate embodiments use other source materials.

[0033] Thin oxide films are deposited at a temperature that is high enough such that, when it is adsorbed to the substrate surface, the vaporized source material reacts with a molecular layer of a second source material or that the vaporized source material becomes adsorbed and reacts with the second source material directed to the substrate surface in the subsequent step. On the other hand, the temperature should be low enough such that thermal breakdown of the source material does not occur or that its significance in terms of the total growth rate of the film is very small. Regarding the above-listed metals, the ALD process is carried out at a temperature range of approximately 200-600°C. Alternate embodiments use other temperature ranges.

[0034] In another embodiment of the NROM memory transistor of the present invention, the high-k dielectric layers can be fabricated using evaporation techniques. Various evaporation techniques are subsequently described for the high dielectric constant materials listed above.

[0035] Very thin films of TiO_2 can be fabricated with electron-gun evaporation from a high purity TiO_2 slug (e.g., 99.9999%) in a vacuum evaporator in the presence of anion beam. In one embodiment, an electron gun is centrally located toward the bottom of the chamber. A heat reflector and a heater surround the substrate holder. Under the substrate holder is an ozonizer ring with many small holes directed to the wafer for uniform distribution of ozone that is needed to compensate for the loss of oxygen in the evaporated TiO_2 film. An ion gun with a fairly large diameter (3 – 4 in. in diameter) is located above the electron gun and argon gas is used to generate Ar ions to bombard the substrate surface uniformly during the film deposition to compact the growing TiO_2 film.

[0036] A two – step process is used in fabricating a high purity HfO_2 film. This method avoids the damage to the silicon surface by Ar ion bombardment, such as that encountered during Hf metal deposition using dc sputtering. A thin Hf film is deposited by simple thermal evaporation. In one embodiment, this is by electron-beam evaporation using a high purity Hf metal slug (e.g., 99.9999%) at a low substrate temperature (e.g., 150° – 200°C). Since there is no plasma and ion bombardment of the substrate (as in the case of sputtering), the original atomically smooth surface of the silicon substrate is maintained. The second step is oxidation to form the desired HfO_2.

[0037] The first step in the deposition of CoTi alloy film is by thermal evaporation. The second step is the low temperature oxidation of the CoTi film at 400°C. Electron beam deposition of the CoTi layer minimizes the effect of contamination during deposition. The CoTi films prepared from an electron gun possess the highest purity because of the high-purity starting material. The purity of zone – refined starting metals can be as high as 99.999%. Higher purity can be obtained in deposited films because of further purification during evaporation.

[0038] A two step process in fabricating a high-purity ZrO_2 film avoids the damage to the silicon surface by Ar ion bombardment. A thin Zr film is deposited by simple thermal evaporation. In one embodiment, this is accomplished by electron beam evaporation using an ultra-high purity Zr metal slug (e.g., 99.9999%) at a low substrate temperature (e.g., 150° – 200°C). Since there is no plasma and ion bombardment of the substrate, the original atomically smooth surface of the silicon substrate is maintained. The second step is the oxidation to form the desired ZrO_2.

[0039] The fabrication of Y_2O_3 and Gd_2O_3 films may be accomplished with a two step process. In one embodiment, an electron gun provides evaporation of high purity (e.g., 99.9999%) Y or Gd metal followed by low-temperature oxidation technology by microwave excitation in a Kr/O_2 mixed high-density plasma at 400 ° C. The method of the present invention avoids damage to the silicon surface by Ar ion bombardment such as that encountered during Y or Gd metal deposition sputtering. A thin film of Y or Gd is deposited by thermal evaporation. In one embodiment, an electron-beam evaporation technique is used with an ultra-high purity Y or Gd metal slug at a low substrate temperature (e.g., 150 ° – 200 ° C). Since there is no plasma or ion bombardment of the substrate, the original atomically smooth surface of the silicon substrate is maintained. The second step is the oxidation to form the desired Y_2O_3 or Gd_2O_3.

[0040] The desired high purity of a PrO_2 film can be accomplished by depositing a thin film by simple thermal evaporation. In one embodiment, this is accomplished by an electron-beam evaporation technique using an ultra-high purity Pr metal slug at a low substrate temperature (e.g., 150 ° – 200 ° C). Since there is no plasma and ion bombardment of the substrate, the original atomically smooth surface of the silicon substrate is maintained. The second step includes the oxidation to form the desired PrO_2.

[0041] The nitridation of the ZrO_2 samples comes after the low-temperature oxygen radical generated in high-density Krypton plasma. The next step is the nitridation of the samples at temperatures > 700 ° C in a rapid thermal annealing setup. Typical heating time of several minutes may be necessary, depending on the sample geometry.

[0042] The formation of a Y-Si-O film may be accomplished in one step by co-evaporation of the metal (Y) and silicon dioxide (SiO_2) without consuming the substrate Si. Under a suitable substrate and two-source arrangement, yttrium is evaporated from one source, and SiO_2 is from another source. A small oxygen leak may help reduce the oxygen deficiency in the film. The evaporation pressure ratio rates can be adjusted easily to adjust the Y-Si-O ratio.

[0043] The prior art fabrication of lanthanum aluminate ($LaAlO_3$) films has been achieved by evaporating single crystal pellets on Si substrates in a vacuum using an electron-beam gun. The evaporation technique of the present invention uses a less expensive form of dry pellets of Al_2O_3 and La_2O_3 using two electron guns with two rate monitors. Each of the two rate monitors is set to control the composition. The composition of the film, however, can be shifted toward the Al_2O_3 or La_2O_3 side depending upon the choice of dielectric constant. After deposition, the wafer is annealed ex situ in an electric furnace at 700 ° C for ten minutes in N_2 ambience. In an alternate embodiment, the wafer is annealed at 800 ° – 900 ° C in RTA for ten to fifteen seconds in N_2 ambience.

[0044] Figure 2 illustrates an energy-band diagram in accordance with the NROM flash transistor of Figure 1. The diagram shows the relationship between E_C, Φ, and the energy difference with respect to the conduction band edge in the high-k trapping layer, E_t.

[0045] Figure 3 illustrates an energy-band diagram in accordance with a write operation in the transistor structure of Figure 1 while Figure 4 is the energy-band diagram for an erase operation. The diagrams show the conduction band edge, E_C, and the valence band edge, E_V. Between E_C and E_V is the band gap where there are no states for electrons. The energy barrier, Φ, is the discontinuity in the conduction bands.

[0046] The high-k tunnel gate dielectric of the present invention reduces the barriers between the substrate and gate insulator and/or between the floating gate and the gate insulator. Figure 5 illustrates a plot of tunneling current dependence on barrier height for various electric fields in accordance with the transistor structure of Figure 1. This plot shows that the tunneling current at a fixed electric field can be increased by orders of magnitude as a result of reducing the barriers.

[0047] In the specific case of Fowler-Nordheim tunneling, the expression that describes the conduction in the insulator is $J = AE^2\exp(-B/E)$ where J is the current density in amps/cm^2, E is the electric field in the insulator in volts/cm and A and B are constants for a particular insulator. The constants depend on the effective mass and the electron barrier energy of the insulator and are scaled with the barrier energy, Φ, as $A \propto (1/\Phi)$ and $B \propto (\Phi)^{3/2}$.

[0048] For the case of the commonly used gate insulator, SiO_2, the equation above renders $A(SiO_2\text{-}Si) = 5.5\text{x}10^{-16}$ amps/volt^2 and $B(SiO_2\text{-}Si) = 7.07\text{x}10^7$ V/cm. If a new barrier of $\Phi = 1.08$ eV is utilized, likely values for A and B can be extrapolated from the above equations. In this case, $A(\Phi=1.08 \text{ eV}) = 1.76\text{x}10^{-15}$ amps/volt^2 and $B(\Phi=1.08 \text{ eV}) = 1.24\text{x}10^7$ V/cm.

[0049] Curves of J versus the barrier energy, F, are shown in Figure 5 for several values of E. For a given tunneling current, lower barriers require lower electric fields. As an example, an SiO_2 barrier of 3.2 eV has an electric field of $6\text{x}10^6$ V/cm while for the same tunneling current, a high-k dielectric with a 1.08 eV barrier requires only an electric field of $7\text{x}10^5$ V/cm. If the thicknesses of the two dielectrics are the same then the voltage required will be about 8.6 times less for the same current. If the high-k dielectric has a dielectric constant of 28, then the equivalent oxide thickness (EOT) will be 7 times less than the actual thickness of the high-k dielectric.

[0050] The NROM memory transistors of the present invention can thus be designed with very small equivalent oxide thicknesses and scaled into the 50 nm dimensions without drain turn-on problems, short-channel effects, and punchthrough. Additionally, retention times will decrease due to more thermal excitation and emission of electrons over the smaller barriers.

[0051] Figure 6 illustrates a functional block diagram of a memory device 600 that can incorporate the flash memory cells of the present invention. The memory device 600 is coupled to a processor 610. The processor 610 may be a microprocessor or some other type of controlling circuitry. The memory device 600 and the processor 610 form part of an electronic system 620. The memory device 600 has been simplified to focus on features of the memory that are helpful in understanding the present invention.

[0052] The memory device includes an array of flash memory cells 630 that can be NROM flash memory cells. The memory array 630 is arranged in banks of rows and columns. The control gates of each row of memory cells is coupled with a wordline while the drain and source connections of the memory cells are coupled to bitlines. As is well known in the art, the connection of the cells to the bitlines depends on whether the array is a NAND architecture or a NOR architecture.

[0053] An address buffer circuit 640 is provided to latch address signals provided on address input connections A0-Ax 642. Address signals are received and decoded by a row decoder 644 and a column decoder 646 to access the memory array 630. It will be appreciated by those skilled in the art, with the benefit of the present description, that the number of address input connections depends on the density and architecture of the memory array 630. That is, the number of addresses increases with both increased memory cell counts and increased bank and block counts.

[0054] The memory device 600 reads data in the memory array 630 by sensing voltage or current changes in the memory array columns using sense/buffer circuitry 650. The sense/buffer circuitry, in one embodiment, is coupled to read and latch a row of data from the memory array 630. Data input and output buffer circuitry 660 is included for bi-directional data communication over a plurality of data connections 662 with the controller 610). Write circuitry 655 is provided to write data to the memory array.

[0055] Control circuitry 670 decodes signals provided on control connections 672 from the processor 610. These signals are used to control the operations on the memory array 630, including data read, data write, and erase operations. The control circuitry 670 may be a state machine, a sequencer, or some other type of controller.

[0056] Since the NROM memory cells of the present invention use a CMOS compatible process, the memory device 600 of Figure 6 may be an embedded device with a CMOS processor.

[0057] The flash memory device illustrated in Figure 6 has been simplified to facilitate a basic understanding of the features of the memory. A more detailed understanding of internal circuitry and functions of flash memories are known to those skilled in the art.

CONCLUSION

[0058] In summary, an NROM cell can use a high-k dielectric as the trapping layer. The high-k dielectric can be fabricated using atomic layer deposition, evaporation, or a combination of the two processes. The high-k dielectric enables smaller write and erase voltages to be used and eliminates drain turn-on problems, short-channel effects, and punchthrough.

[0059] The NROM flash memory cells of the present invention may be NAND-type cells, NOR-type cells, or any other type of array architecture.

[0060] Although specific embodiments have been illustrated and described herein, it will be appreciated by those of ordinary skill in the art that any arrangement that is calculated to achieve the same purpose may be substituted for the specific embodiments shown. Many adaptations of the invention will be apparent to those of ordinary skill in the art. Accordingly, this application is intended to cover any adaptations or variations of the invention. It is manifestly intended that this invention be limited only by the following claims and equivalents thereof.